Examining Item Position Effect on Reliability Coefficients of Multiple Choice Physics Achievement Test

Fagbenro W. Ayoola, Abdullahi Ibrahim

Department of Science Education, Federal University Wukari, Nigeria.

DOI: https://dx.doi.org/10.47772/IJRISS.2024.806174

Received: 05 June 2024; Revised: 23 June 2024; Accepted: 27 June 2024; Published: 16 July 2024

ABSTRACT

This study explored the effect of changes in item sequence on test reliability coefficients (test-retest and Kuder Richardson 20) and student's achievement in multiple-choice physics tests in Senior Secondary School II in Taraba State. The study adopted repeated measures two-group within-subject experimental research design. The research collected data in order to answer three research questions and test one hypothesis. The sample comprised 450 senior secondary II Physics students (male = 303 and female = 147) drawn from population of Physics students in Taraba State. Multi stage sampling technique was employed to randomly select twelve schools from three Local Government Areas of the three Senatorial Districts of Taraba State, and an intact arm of SS II from each of the sampled schools was used. Two parallel 40-items Physics Achievement Test developed by the researcher were used for data collection. The resulting data were collated and analysed using descriptive statistics and t-test. Results showed that reliability coefficients obtained from the two formats were quite comparable. However, the finding revealed that there is significant difference between the students mean achievement score in the format A and Format B of the physics achievement test (t = 4.409, df = 898, $p < 0.05$, two-tailed). Sequels to the findings, it is concluded that changes in item sequence has no effect on the reliability coefficients, but students will perform better in physics achievement test if the test items are arranged randomly than in descending order of difficulty.

Keywords: Item sequence, physics achievement test, reliability coefficient, test-retest, Kuder-Richardson 20 (K-R 20).

INTRODUCTION

Testing and assessment of abilities in learners is one of the main focuses of educational measurement. The most important reason for carrying out assessment is to verify students' knowledge and to find out "how" they know it. Since teachers can't see into students' heads, and knowledge is imperceptible, observation of student's behaviour or performance is one of the methods of determining what students' know (Mislevy, 2003). Measurement experts always talk about measures of ability, which could either be measures of achievements or measure of ability. According to Baker and Kim (2004), educational tests try to measure several or one hypothetical construct that is usually unobservable, referred to as latent trait. Intelligence and arithmetic ability are examples of latent traits. Consistency in measurement or reliability of test is required in measuring latent traits. Conceptually, reliability is the fraction of the whole variance in an obtained score that is due to true variance as opposed to error variance (Cohen & Swerdlik, 2010; Schmidt & Embretson, 2013). In reality, reliability has to do with consistency of test and manifest in different forms. The degree of correlation between pairs of scores from the same individuals on two separate administrations of the same test is taken into account when assessing test-retest consistency. Parallel and alternate-form reliability refers to the reliability estimates that compare two parallel (means and variances of observed test scores are equal) or alternate test forms to ascertain how much they correlate with one another. The degree of correlation between items on a test, whether by comparing the two halves of a single test (split-half reliability) or by comparing each item with every other item on a test or scale (inter-item correlation, item total correlation, Cronbach's alpha), is known as internal consistency. In order for a measure to be considered psychometrically sound, it is crucial to assess the instrument's stability.

According to Duke (1979), a test's consistency affects its standard error of measurement (SEM). SEM is a measure of the predicted variability for observed scores under the assumption that the true score remains constant. Because few tests demonstrate full reliability, there is variability in observed scores; therefore, the lower the SEM, the higher the reliability. A lower SEM necessitates significant effort in mistake minimization. Lower test precision and dubious test validity are associated with greater SEMs. SEM is also estimated using internal consistency reliability (Slick, 2004). Using SEM, one can obtain the range of scores that could include the true value, often known as the confidence interval (Cohen & Swerdlik, 2010). For instance, if a student received an observed standard score of 32 on an achievement test with a test's SEM of 2, we can be 96% positive that her real score falls between 28 and 36 (62 ± 2 SEM) or 68% confident that it falls between 30 and 34 (62 ± 2 SEM).

Aside from the SEM, there are additional factors that can cause errors in test results. These include examiner factors (such as test administration and rapport-building skills, scoring and interpretation errors), examinee factors (such as guesswork, fatigue, lack of motivation, and reactivity to the testing situation), and environmental factors (such as temperature, lighting, and noise level in the room). It is unlikely that an unmotivated examinee's ability on an achievement test in physics given in a noisy environment will be accurately reflected. It is assumed that the test taker is working hard in a setting that minimizes irrelevant factors (i.e., things unrelated to physics). Such a mistake may reduce a test's reliability and cause bias in the results. Differential Item Functioning (DIF), in which the likelihood of accepting an item is higher for one group than the other, across different characteristic levels, is another type of bias in test results (Swaminathan & Rogers, 1990). In other words, even though two individuals from distinct groups possess the same latent characteristic level, their likelihood of correctly answering a particular item varies. Test bias and DIF are connected to the idea of measurement equivalency. When there are consistent relationships between latent traits and observed test scores across many populations, measurement equivalency occurs (Drasgow, 1984). Testing tests for DIF and measurement equivalency is crucial to guaranteeing that test results are appropriately interpreted across samples.

SEM and reliability estimates are associated with the classical test theory (CTT), which is a collection of guidelines for assessing how well tests estimate unobservable variables of interest (DeVellis, 2006; Gulliksen, 1950; Lord & Novick, 1968). In addition to true score variance and error variance, in CTT, an observed score variance and an observed test score are functions of a true and an error score (Spearman, 1907, 1913). The assumptions that underpin CTT are listed by Schmidt and Embretson (2013) and Zickar and Broadfoot (2008). Since errors are random, the initial assumption is that there is no correlation between true and error scores. The second premise holds that since errors are random and caused by a variety of factors, a normal distribution of errors should be anticipated. Consequently, for every test subject in the population and throughout replications, the mean error score is zero. Third, there is no correlation between mistake scores and other test results or scores on parallel exams.

Item feature effects have been shown in recent MC research in science education to have a significant impact on the extraction and assessment of student knowledge. For instance, higher poise in response accuracy in physics education and better performance on MC assessments (Caleon & Subramaniam, 2010) and chemistry education (Rodrigues, Taylor, Cameron, Syme-Smith, & Fortuna, 2010) were correlated with prior knowledge of the behavioral trait to be tested. One popular technique to increase the reliability of MC test results is to rearrange the items' positions or locations during the test (Bulut, Quo & Gierl, 2017). As a result, issues like people remembering questions or mimicking other test takers' responses can be resolved

(Bulut, 2015). Exam malpractice was addressed, which could have affected the test's psychometric qualities. Nevertheless, item location effects resulted from this strategy (Bulut, 2015). In many testing situations, the impact of item position on an individual's talents is disregarded. If it happens, it is thought to apply to every individual and every item (Hahne, 2008; Albano, 2013). The validity of test score interpretations is threatened in practice by the fact that individual test results can differ based on item position (Albano, 2013; Trendtel & Robitzsch, 2018). Differential item functioning (DIF) can result from the arrangement of items in test forms made by manipulating item placements (Debeer & Janssen, 2013; Erdem, 2015; Balta & Omur Sunbul, 2017; Akayleh, 2018). While some studies (Hartig & Buchholz, 2012; Debeer & Janssen, 2013; Ollennu & Etsey, 2015). The West African Examinations Council [WAEC], 1993) suggest that examinees' achievement is impacted by item position, other studies (Perlini et al., 1988; Tal et al., 2008; DoğanGül & Çokluk Bökeoğlu, 2018) have come to the opposite conclusion. It was found in several studies (Meyers et al., 2009; Debeer & Janssen, 2013; Hecht et al., 2015; Doğan Gül & Çokluk Bökeoğlu, 2018) that item position led to bias in item parameter estimations. While the majority of studies on IP effects use the Classical Test Theory (CTT) as their foundation, there are also studies that use the Item Response Theory (IRT) framework (Hahne, 2008; Hohensinn et al., 2008; Debeer & Janssen, 2013; Qian, 2014; Weirich et al., 2014).

Effect of Item Position on Achievement

Any influence or interpretation that an item may get purely as a result of its relationship with other items in a specific test is called item context effect. The "influence" and "interpretation" refer to the effect acquired by the item, while the relationship has to do with cause or source of the effect, the other items in the test or, its context. The other items in the test or context may be described in many ways. The other items may be designed to measure different content at different level of cognitive understanding or different in difficulty or type.

For example, items may be different in content assessed, task type, response type, in the number of options. Overall, items may be less difficult or more difficult than one another. The amount of effort and time needed by the items may be different, and there may simply be a smaller or larger total number of items. More reasoning or compound constructed responses may be needed by complex items, whereas simpler items may require only a few seconds to complete.

Finaly, items may be different in discriminating ability and in quality. So, as any one of these features changes, the context of an item within the test forms changes as well. Context is usually defined in reference to the portion of the test form which the test takers' have responded to and which will have the most effect on their response to the target item. That is, the reference to the items before a target item. The items which follow a target item may also contribute to the definition of context and position. Such as when a test taker is free to skip and then come back to item or go back to review previous responses. In speeded tests, depending on how much testing time is left, the number of items left in the test may affect performance on the target item.

In this study, context effect involved a simple design, the test forms contain the same items and the only difference is their arrangement in the forms. Examinee group taking each form will be assumed to be comparable in ability since forms contained the same items, since it is done through random assignment. So, any differences across the score distribution are attributable to the rearrangement of items in the different forms. The statistic that will be used in the arrangement of the test items is the item difficulty. The percentage of students who correctly answered a question is called item difficulty, or p-value. From this definition, items answered correctly by most examinees will have high p-value and is therefore an easy item. Conversely, a low p-value will be determined by using group similar to the targeted group.

Opara & Uwah (2017) investigated the effect of test item arrangement on performance in Mathematics among Junior Secondary School Students in Obio-Akpor L.G.A of Rivers State. The study adopted the quasi-experimental research design. A sample of one hundred (100) Junior Secondary School II Students drawn from the population of 6,777 students, from four public schools in the area, using the simple random and non-proportionate sampling techniques. The sample included three experimental group labelled 'A, B and C" and the control group labelled "D". Mathematics Achievement Test" which had four types A, B, C and D was used. Type A was arranged in ascending order of difficulty. Type B was arranged in descending order of difficulty while type C was arranged based on order of topic presentation in the class. On the other hand, type D was arranged in no particular order. Validity of the instruments was determined using table of specification (TOS) while a general reliability index of 0.96 was determined using Kuder Richardson formular 20 (KR20). Mean, standard deviation as well as t-test analysis were used to compare the mean of each of the group against the control group and test. The findings of the study were that item arrangement based on ascending order of difficulty has a positive and significant effect on students' performance in mathematics at 0.05 alpha level respectively while item arrangement based on descending order has a positive but insignificant effect on student' performance in mathematics. Finally, item arrangement based on no particular order of difficulty has a positive and significant effect on students' performance. It was recommended among others that classroom teachers, test constructors and professional examination bodies should endeavour to arrange items from simple to complex in order to boost students' morale.

Mohammadi (2014) examined the effect of test item order in relation to course content sequence on two hundred and fifty nine (259) nursing students in Novel Drug Delivery Research Center, Kermanshah, Iran. The instrument consisted of four option multiple-choice items arranged in three formats. In the first format, items were arranged according to the course plan, the second format items were arranged opposite to the sequence of course plan, and the third format was randomly sequenced. The subjects were randomly selected and assigned to group and treatment. The data was analysed using SPSS (11.5) package. Descriptive statistics and Analysis of Variance (ANOVA) tests were use. It was found that the sequencing has a significant effect on the achievement of the students (F (2,256) = 0.565, p = 0.566).

Kristin & Allison (2013) explored whether the ordering of the questions make a difference as to how students perform in a test. The test were either constructed with easiest items first and hardest item listed last (A), or the other way with the hardest item first on the test and the easiest item placed at the end of the test (B), but each test had the exam same items. Easy items were defined as those that involved one-step calculations without the application of the concept in a word problem. Also easy questions were those that required simple factual recall. Hard questions were defined as those that required multiple steps, especially those questions embedded in word problem and application. Tests were handed out randomly to five hundred and ninety-four (594) mathematics majors and non-majors students. Data were collected over multiple semesters with several different classes. Analysis was performed using SAS. It was found that most of the mathematics students who were examined, the ordering of the questions on a test did not impact performance

Naibi Louisa (2013) investigated the effects of two types of item arrangement formats (ascending and specified mixed order), on two types of test reliability coefficients (test-retest and Kuder Richardson 20). A repeated measures two-group within-subject design was used, with a sample of four hundred and eighty (480) senior secondary school mathematics students from three Local Government Areas in Bayelsa State, Nigeria. A 40-item Mathematics Achievement Test was used to collect data, which was analyzed using the z-test and the t-formula for testing the significance of reliability coefficients. The study finds significant effect on test scores, with the specified mixed order format having significantly higher scores, but found no effect on either of the reliability coefficients. It is recommended that item arrangement be strictly considered when constructing tests in schools and for standardized examinations.

When a factor other than ability—in this case, item position—affects a student's likelihood of answering a question correctly or incorrectly, that question is biased in the assessment context. The examined research findings about how item arrangement affects students' achievement are conflicting. Additionally, a number of studies have demonstrated that item arrangement has an impact on test scores and that variables have been linked to test scores. Any factor influencing test score reliability jeopardizes objectivity, accuracy, precision, external validity, and overall validity since test results determine the psychometric criteria crucial to measurement strategy. Studying this effect and offering strategies for evaluating, regulating, and reducing it is crucial. The degree to which item position in a multiple-choice test would affect test takers' achievement is still contentious. There is no unanimous agreement among the researchers as revealed by the literature reviewed. The inability of the researchers to reach agreement on the potential risk of administering tests with different arrangements of the same item to students provided the motivation for the present study. The purpose of this research was to investigate the consequence of changing item arrangement on the students' achievement on multiple – choice physics questions. More specifically, the present study sought to: compare students' mean scores of two different test formats in physics and determine the responses to the two formats of physics achievement tests affect the value of reliability coefficient. The present study postulates that there exists no statistically significant variation in the average test scores, test-retest reliability coefficient, and K-R 20 reliability coefficient of an accomplishment test in physics between the formats of ascending order items and random mixed order items.

METHOD

The repeated measures two-group within-subject experimental research design was employed in the study. Because the model evaluates a subject's response to each treatment, the subjects in this design act as their own controls (Kerlinger & Lee, 2000). In this situation, the variable being manipulated is the item position. The treatment is the variable being manipulated whose effect is being investigated. Two groups of subjects were chosen, and treatments were assigned to the groups at random. On the Physics Achievement test, the first group completed format A answers first, while the second group completed format B answers first. During the second administration, the first group provided format B answers, whereas the second group provided format A answers.

Population

Every student enrolled in Taraba State's Senior Secondary School II in Nigeria was the study's target group. Taraba State's Senior Secondary School II is anticipated to have 9,528 pupils enrolled overall. Due to the fact that SS II students are not prepared for any external examinations, they were selected as study participants. It is further expected that the students have studied enough physics (electricity) to be able to answer any questions the researcher poses.

Sample and Sampling Techniques

A multistage random sampling technique was used as the sampling method. Initially, Taraba State's sixteen (16) local government areas were divided into three senatorial districts through stratification. A single local government unit was chosen at random from each of the three senatorial districts. Private and public secondary schools were distinguished in each of the three local government districts that were chosen at random. From each local government region, two public and two private secondary schools were chosen at random. The selected arms in the schools were an intact class. The sample size was four hundred and five (405). The information is contained in table1 and 2.

Table 1: Distribution of Senior Secondary School II Science Students in Taraba State by Council Area and Sex

S/N	Local Council Areas	No. of School	Senior Secondary School II Science Students		
			Male	Female	Total
1	IBI	15	175	125	300

2	WUKARI	29	450	270	720
3	DONGA	15	181	119	300
4	USSA	14	196	154	350
5	TAKUM	26	465	445	910
6	GASSOL	16	170	150	320
7	BALI	15	175	140	315
8	SARDAUNA	15	165	152	317
9	KURUMI	14	245	155	400
10	GASHAKA	24	417	299	716
11	LAU	10	102	88	190
12	ZING	12	225	195	420
13	ARDO-KOLA	12	231	189	420
14	KARIN LAMIDO	45	1,148	652	1,800
15	YORRO	12	250	200	450
16	JALINGO	35	985	615	1,600
	TOTAL	**309**	**5580**	**3,948**	**9528**

Table 2: Sample Frame for Effect of Item Sequence on Physics Achievement

S/N	Selected Local Government Area	No of School	Sample of SSII	Sample of SSII by Sex	
				Male	**Female**
1	Wukari	4	154	108	46
2	Jalingo	4	193	134	59
3	Bali	4	103	61	42
Total		**12**	**450**	**303**	**147**

Instrument

The Physics Achievement Test was the instrument used in this study to collect data. There were sixty items in the first draft of the PAT. It was put together by the researcher. Nonetheless, consideration was given to the physics curriculum created by the Federal Ministry of Education in Abuja, Nigeria, as well as the Senior Secondary Certificate Examination syllabus created by West Africa Examination Council and Nigeria Examination Council. The contents of the senior secondary one and senior secondary two physics curricula served as the basis for the development of the items. Furthermore, the items were written using the NECO and WAEC patterns. In other words, the four response options for each item were A, B, C, and D. The items covered one main theme in physics, this is electricity.

First and foremost, an examination of the curricula in every school included in the sample revealed that every physics instructor had taught electricity, which led to the choice to create items centred around electricity. Two, students need to become proficient in a large number of formulas and equations related to electricity. The Physics Chief Examiners (WAEC, 2012) have observed that a lot of candidates struggle with using formulae and equations in exam items. The purpose of the test blue print, which is presented in Table 5, was for ensuring the test’s content validity. Due to the pupils’ ages and the need to minimize boredom, the thought processes were restricted to knowledge, comprehension, and application. In addition, a panel of knowledgeable specialists in education evaluation and physics education was consulted to see whether the items were appropriate. The instrument has logical validity index of 0.77.

To ascertain the difficulty index, 150 students were initially administered a draft copy of the PAT, which had 60 items. The instrument was given to the students for a total of sixty minutes. The test took the students, on average, fifty minutes to complete. CTT was used for item analysis in order to choose the final items. Items that didn't meet the requirements were eliminated based on the difficulty index criteria, which are 0.30 $\geq p \geq$ 0.80. The Physics Achievement Test comes in two formats: format A has 40 items that are placed at random, while format B has 40 items that are arranged in a "hard to easy" order according to difficulty.

Data Analysis and Results

The result of the statistical analysis are presented and discussed in this chapter. The order of presentation follows the order of the research questions stated in chapter one.

Data Presentation

The results of data analysis using descriptive statistics and t-tests are presented in tables. Tests of hypotheses and results of acceptance or rejection of null hypotheses and research questions are analyzed using SPSS and the results presented in tables.

Research Questions

1. How do the students' means scores from the two different formats compare as equivalent test?
2. To what extent is the reliability coefficient using test-retest method affected by different test formats.
3. To what extent is the reliability coefficient using K-R20 method affected by different test formats.

Hypothesis

There is no significant difference in the students' mean scores when the item sequence of multiple-choice physics test is changed.

Data Presentation

Test of hypothesis and results of acceptance or rejection of null hypothesis and research questions are analyzed using SPSS and the results presented in tables.

Research Question 1: How do students' mean scores from the two different formats compare as equivalent tests?

The Physics Achievement Scores were grouped into two categories based on the format. Four hundred and fifty subjects were involved in the research. The data was analysed using descriptive statistics. The summary of the descriptive statistics is presented in table 3 below.

Table 3: Summary of the descriptive statistics of Achievement in Physics Scores

	Question format	N	Mean	Std. Deviation	Std. Error Mean
Physics Achievement Test Score	Format A	450	22.2867	5.19466	0.24488
	Format B	450	20.7000	5.59480	0.26374

Table 3 shows that the mean score of the students from Physics Achievement Test Score Format A is 22.29 with standard deviation of 5.2, and that of Format B is 20.70 with standard deviation of 5.6. The mean difference is 1.5867. There is a difference between the students' mean scores of the two different formats.

Research Question 4: To what extent is the reliability coefficient using test-retest method affected by different test formats.

The reliability coefficient was estimated using test-retest method for the two different formats of physics achievement question. Summary is presented in table 4 below.

Table 4: Reliability Coefficients Using Test-retest

	Question format	N	$R_{test\text{-}retest}$
Physics Achievement Test	Format A	450	0.851
	Format B	450	0.832

From the table 4 above, the reliability coefficient of Physics Achievement Test format A is 0.851, and that of format B is 0.832. Though the two Physics Achievement Tests contain the same items, their reliability coefficients are not the same.

In a physics achievement format A test (40-item) with a test-retest reliability coefficient of 0.851 and a standard deviation of 5.19466 would have a Standard Error of Measurement of 2. If a student had a score of 34, her scores would be expected to fall between 32 and 34 (if the test is taken repeatedly). To be 95% confident in the test scores, we would look at the interval of 34 ± 1.96 (2).

A physics achievement format B test (40-item) with a test-retest reliability coefficient of 0.831 and a standard deviation of 5.59480 would have a Standard Error of Measurement of 2.30. If a student had a score of 34, her scores would be expected to fall between 31.70 and 34.3 (if the test is taken repeatedly). To be 95% confident in the test scores, we would look at the interval of 34 ± 1.96 (2.3). It follows that the reliability coefficient is not affected by position of items in the test formats.

Research Question 3: To what extent is the reliability coefficient using K-R20 method affected by different test formats.

The reliability coefficient was estimated using K-R20 method for the two different formats of physics achievement test. Summary is presented in table 5 below.

Table 5: Reliability Coefficients Using K-R20

	Question format	N	$R_{K\text{-}R20}$
Physics Achievement Test	Format A	450	0.847
	Format B	450	0.830

From the table above, the reliability coefficient of Physics Achievement Test format A is 0.847, and that of format B is 0.830. Though the two Physics Achievement Tests contain the same items, their reliability coefficients are not the same.

A physics achievement format A test (40-item) with a test-retest reliability coefficient of 0.847 and a standard deviation of 5.19466 would have a Standard Error of Measurement of 2.03. If a student had a score of 34, her scores would be expected to fall between 31.97 and 34.03 (if the test is taken repeatedly). To be

95% confident in the test scores, we would look at the interval of 34 ± 1.96 (2.03).

A physics achievement format B test (40-item) with a test-retest reliability coefficient of 0.830 and a standard deviation of 5.59480 would have a Standard Error of Measurement of 2.306. If a student had a score of 34, her scores would be expected to fall between 31.694 and 34.306 (if the test is taken repeatedly). To be 95% confident in the test scores, we would look at the interval of 34 ± 1.96 (2.306). It follows that the reliability coefficient is not affected by position of items in the test formats.

Hypothesis 1: There is no significant difference in the students' mean scores when the item sequence of multiple-choice physics test is changed.

The scores were analysed using descriptive statistics and student's t test. The descriptive statistics and the student's t table are presented in table 6 and 7 respectively. As shown in table 6, the mean achievement score in Physics Achievement Test Format A ($\bar{X}$ = 22.2867, S.D = 5.19) is more than the mean achievement score in Physics Achievement Test Format B ($\bar{X}$ = 20.7000, S.D = 5.59). The mean difference between the two conditions was 1.5867.

Table 6: Descriptive of Achievement Scores by Format

	Question format	N	Mean	Std. Deviation	Std. Error Mean
Physics Achievement Test Score	Format A	450	22.2867	5.19466	0.24488
	Format B	450	20.7000	5.59480	0.26374

Table 7: The Student's t Table

	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Physics Achievement Test Score	4.409	898	.000	1.58667	0.35990

From the table6 above, an independent t-test showed that the difference between conditions was significant (t = 4.409, df = 898, p = 0.000, two-tailed).

The null hypothesis of no significant difference between the mean scores of the two conditions, when items are arranged randomly and when they are arranged in descending order of difficulty, is rejected. This implies that there is a significant difference in the mean achievement score when the item sequence of the multiple-choice physics achievement test is changed.

DISCUSSION OF FINDINGS

The students mean achievement score in the format A of the Physics Achievement test is higher than their mean achievement score in Format B of the Physics Achievement test, and the different is statistically significant. This implies that the students will achieved better if the items are arranged randomly than in the descending order of item difficulty. This might be due to time spent on difficult items before arriving at the easy items resulting in the inability of the students to finish within the allotted time. Therefore, the hypothesis of no significant difference in the mean score between formats is rejected. It follows that the manner in which the items on physics achievement test are arranged affects students' achievement. This is in agreement with Opara & Uwah (2017) that investigated the effect of test item arrangement on performance in Mathematics among Junior Secondary School Students in Obio-Akpor L.G.A of Rivers State, and recommended among others that classroom teachers, test constructors and professional examination bodies should endeavour to arrange items from simple to complex in order to boost students' morale; Mohammadi (2014) that examined the effect of test item order in relation to course content sequence on two hundred and fifty nine (259) nursing students in Novel Drug Delivery Research Center, Kermanshah, Iran, and found that the sequencing has a significant effect on the achievement of the students (F (2,256) = 0.565, p = 0.566) ; Naibi Louisa (2013) investigated the effects of two types of item arrangement formats (ascending and specified mixed order), on two types of test reliability coefficients (test-retest and Kuder Richardson 20) and found significant effect on test scores, with the specified mixed order format having significantly higher scores and Kamal (2011) that investigated the effect of sequencing items of a grammar test from easy to difficult versus difficult to easy on achievement of Iranian foreign language learners of English and found that sequencing has a significant effect on the achievement of the students . This result is not in agreement with Russell (2003) who administered students in two sections of an Advertising course and one section of a Sales Management course as well as students in three sections of management courses with three different multiple choice exams over one semester and concluded that the differences in student performance across the three exam versions were not significant; Kristin & Allison (2013) explored whether the ordering of the questions make a difference as to how students perform in a test and found that the ordering of the questions on a test did not impact performance.

The reliability coefficients of the two test formats were estimated using test-retest and K-R20 method. The reliability coefficients values using test-retest and K-R20 methods are comparable. The different positions of the items in the two formats have effect on the reliability coefficient estimate. This result is in agreement with Naibi Louisa (2013) that investigated the effects of two types of item arrangement formats (ascending and specified mixed order), on two types of test reliability coefficients (test-retest and Kuder Richardson 20) and found no effect on either of the reliability coefficients

CONCLUSION AND RECOMMENDATIONS

This study investigated the effects of changes in item sequence reliability coefficient of multiple-choice physics tests in Taraba state of Nigeria. The findings revealed that arrangement of items of Physics achievement test is associated with Achievement in Physics. Arranging the items randomly make students achieve better than when they are arrange in descending order of difficulty but has no effect on reliability coefficient

Sequels to the findings, educationist should caution institutions involve in high stake test that presenting items that made up an achievement test in physics in different arrangement to students bias the test and violate the principle of common metric.

REFERENCES

1. Akayleh, A. S. A. (2018). *Precision of the estimations for some methods of the CTT and IRT as a base to display the differential item functions on the different item ordered test formats.* https://bit.ly/3aJeFKx
2. Balta, E., & Omur Sunbul, S. (2017). An investigation of ordering test items differently depending on their difficulty level by differential item functioning. *Eurasian Journal of Educational Research, 72*, 23-42. https://doi.org/doi:10.14689/ejer.2017.72.2
3. Baker, F. N., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques.* Second Edition. New York, NY: Marcel Dekker.
4. Bulut, O. (2015). An empirical analysis of gender-based DIF due to test booklet effect. *European Journal of Research on Education, 3*(1), 7-16. https://bit.ly/3cKkhqf
5. Bulut, O., Quo, Q., & Gierl, M. J. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large-scale Assessments in Education, 5*(1), 8. http://doi.org/10.1186/s40536-017-0042-x
6. Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment: An introduction to tests and measurement.* New York, NY: McGraw-Hill.

7. Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185. https://ppw.kuleuven .be/okp/_pdf/DeBeer2013MIPEW.pdf
8. DeVellis, R. F. (2006). Classical test theory. *Medical care, 44*(11), S50-S59.
9. Doğan Gül, Ç., & Çokluk Bökeoğlu, Ö. (2018). The comparison of academic success of students with low and high anxiety levels in tests varying in item difficulty. *Inonu University Journal of the Faculty of Education, 19*(3), 252-265. https://doi.org/10.17679 /inuefd.341477
10. Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95,* 134-135.
11. Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335-337.
12. Erdem, B. (2015). *Ortaöğretime geçişte kullanılan ortak sınavların değişen madde fonksiyonu açısından kitapçık türlerine göre farklı yöntemlerle incelenmesi* [Investigation of Common Exams Used in Transition to High Schools in Terms of Differential Item Functioning Regarding Booklet Types with Different Methods] [Unpublished master dissertation]. Hacettepe University. Ankara.
13. Gulliksen, H. (1950). *Theory of mental tests.* New York, NY: Wiley. 52
14. Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379–390. https://bit.ly/3aHHyGD
15. Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418- 431. https://core.ac.uk/download/pdf/25705605.pdf
16. Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement, 75* (6), 1021-1044. https://doi.org/10.1177/0013164415573311
17. Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the linear logistic test model. *Psychology Science Quarterly*, 50, 391-402. https://bit.ly/39Sb9xY
18. Kamal Heidari S. (2011). Item Sequence on Test Performance: Easy Items First? *Language Testing in Asia*, 1(3), 46-59.
19. Kristin Kennedy & Allison G. Butler (2013). Changing the Order of Mathematics Test Items: Helping or Hindering Student Performance? *Journal of Humanistic Mathematics* , 3(1), 20-32.
20. Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison Weasley.
21. Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT based common item equating design. *Applied Measurement in Education, 22*(1), 38-60. https://doi.org/10.1080/08957340802558342
22. Mislevy, R. J. (2003). *A Brief Introduction to Evidence-Centered Design* (Technical). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).
23. Mohammadi Farani (2014). A. Test item order in relation to course plan content sequence: Does it affect nursing students' grades in pharmacology exam? *Educ Res Med Sci.*; 3(2), 26-28.
24. Naibi, Louisa (2013). Effect of Item Arrangement on Test Reliability Coefficients: Implications for Testing. *Journal of Research in Education and Society,* 4(3), 54-58.
25. Ollennu, S. N. N., & Etsey, Y. K. A. (2015). The impact of item position in multiple-choice test on student performance at the basic education certificate examination (BECE) level. *Universal Journal of Educational Research, 3*(10), 718-723. https://doi.org/10.13 189/ujer.2015.031009
26. Opara Ijeoma M. & Uwah Idongesit V. (2017). Effect of Test Item Arrangement on Performance in Mathematics among Junior Secondary School Students in Obio/Akpor Local Government Area of Rivers State Nigeria. *British Journal of Education*, 5(8), 1-9.
27. Perlini, A. H., Lind, D. L., & Zumbo, B. D. (1998). Context effects on examinations: The effects of time, item order and item difficulty. *Canadian Psychology/Psychologie Canadienne, 39*(4), 299-307. https://doi.org/10.1037/h0086821

28. Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38(*7), 518-534. https://doi.org/10.1177/014662161453431
29. Russell, Michael, Michael J. Fisher, Carol M. Fisher, & Kathleen Premo (2003). Exam Question Sequencing Effects on Marketing and Management Sciences Student Performance. *Journal for Advancement of Marketing Education,* 3, 1–10.
30. Schmidt, K. M., & Embretson, S. E. (2013). Item response theory and measuring abilities. In J. A. Schinka and W. F. Velicer (Eds.), Research Methods in Psychology (2nd ed.). Volume 2 of Handbook of Psychology (I. B. Weiner, Editor-in-Chief).
31. Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *The American Journal of Psychology*, 161-169.
32. Spearman, C. (1913). Correlations of sums or differences. *British Journal of Psychology, 1904-1920, 5* (4), 417-426.
33. Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.
34. Tal, I. R., Akers, K. G. & Hodge, K. G. (2008). Effect of Paper color and question order on exam performance. Teaching of Psychology, *35*(1), *26-28*. https://doi.org/10.1080/0098 6280701818482
35. The West African Examinations Council [WAEC] (1993). *The effects of item position on performance in multiple choice tests.* Research Report, Research Division, WAEC, Lagos.
36. Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling, 60*(2), 241-263. https://bit.ly/3cQWkh5
37. Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38,* 535-548. https://doi.org/10.1177/0146621614534955
38. Zickar, M. J., & Broadfoot, A. A. (2009). The partial revival of a dead horse? Comparing classical test theory and item response theory. In C. E. Lance, & R. J. Vandenberg. (Eds.). Statistical and methodological myths and urban legends: Doctrine, verity and fable in the organizational and social sciences (pp. 37-59). New York, NY: Routledge/Taylor & Francis Group.